

# Letter to the Editor

## Unveiling the Molecular Clock in the Presence of Recombination

David Posada<sup>1</sup>

Department of Zoology, Brigham Young University

In a recent letter in *Molecular Biology and Evolution*, Schierup and Hein (2000a) showed that the likelihood ratio test (LRT) of the molecular clock (Felsenstein 1981) “wrongly” rejects the clock hypothesis when recombination has occurred. However, this result should not be taken as a failure of the LRT. Because in the presence of recombination often there is not one single tree describing the history of the sequences, but several, the LRT is correctly rejecting the actual null hypothesis tested, that is, that the data are evolving under a clock on one single tree.

To appropriately test the clock hypothesis in the presence of recombination, we need to use a test independent of tree topology. Muse and Weir (1992) proposed a triplet likelihood ratio test to test for equality of evolutionary rates for two species at a time using a third species as an outgroup (to avoid confusion, I will call this test the relative-rate test [RRT]). The RRT is therefore independent of topology and might be used for potentially recombinant sequences if an outgroup is selected which did not recombine with the ingroup. Here, the performance of the RRT with recombinant data is presented.

Recombinant alignments were simulated using the coalescent with recombination (Hudson 1983). A program written in C for this purpose is available from the author. Alignments of 11 sequences (10 recombining ingroup and one nonrecombining outgroup) with 1,000 nt were evolved with a molecular clock under the Jukes-Cantor (JC) model of evolution (Jukes and Cantor 1969). For each level of recombination and diversity, 1,000 replicates were generated. A maximum-likelihood (ML) tree was estimated for each simulated data set under the JC+ $\Gamma$  model of evolution in PAUP\* (Swofford 1998) without assuming a clock. The  $\Gamma$  distribution for rate variation among sites (Yang 1993) was included because it is known that recombination introduces such rate heterogeneity (Schierup and Hein 2000b), and the likelihood increases significantly when this variation is accounted for. For the ML tree, the likelihood under the unconstrained model, where each lineage is allowed to have its own rate (alternative hypothesis), was compared with the likelihood obtained when the molecular clock was enforced (null hypothesis). If the data are evolving under a clock, the difference in likelihood between these two models should be close to zero. To establish statis-

tical significance, twice the difference in likelihood is assumed to be distributed as a  $\chi^2$  with  $n - 2$  degrees of freedom, where  $n$  is the number of sequences (Felsenstein 1981). The likelihoods for the LRTs were calculated in PAUP\*, while the RRTs were performed with HYPHY (Kosakovsky and Muse 2000). In the latter case, multiple tests were calculated for each data set, and the Bonferroni correction was applied to avoid an increase in false positives. If any of the pairwise tests for a given data set were significant, the RRT was considered to reject the clock hypothesis for that data set.

As Schierup and Hein (2000b) previously demonstrated, the LRT rejected the molecular-clock hypothesis even with low levels of recombination (table 1). Increasing divergence made the LRT more prone to reject the molecular clock. However, when the RRT was used, results were very different, and increasing levels of recombination did not affect the rejection levels (table 1). The RRT rejected (after the Bonferroni correction) the clock hypothesis less than 5% of the time (around 2%–3%), so it is a conservative test. This could be due to the lack of power of the relative ratio tests under some conditions (Bromham et al. 2000), but it could also be due to the conservative Bonferroni correction (Rice 1989). More powerful Bonferroni corrections exist that could be easily applied to single data sets (Hochberg 1988).

To illustrate the application of the RRT, four empirical data sets were analyzed (table 2). These data sets are available from the author on request. The first data set was that of the *argF* gene from *Neisseria*, for which intragenic recombination has been suggested (Zhou and Spratt 1992; Grassly and Holmes 1997). For this data set, the LRT rejected the molecular clock, while there were no significant RRTs, which suggests that a clock might be operating in this gene, with the LRT of the clock being significant due to the presence of recombination. The *ND4* gene from the subfamily of lizards *Gymnophthalmi* (Pellegrino et al., personal communication) is mitochondrial and therefore putatively free of recombination. In such a case, both the LRT and the RRT should give us the same answer. Indeed, this was what we observed: both tests failed to reject the molecular clock, which suggests that the evolution of this gene is clocklike and that recombination has not occurred. The third data set was an alignment from the Los Alamos HIV database of HIV-1 *env* sequences from different subtypes within the M group, including known recombinants. Because of the intensive selective pressures at this gene, the data might not conform to a molecular clock. In this case, both tests rejected the molecular clock. In the case of the LRT, the *P* value could be further influenced (downwards) by the likely occurrence of recombination. In the case of the *COI* gene from five vertebrate species (Cunningham 1997), recombination is

<sup>1</sup> Present address: Variagenics, Inc., Cambridge, Massachusetts.

Key words: recombination, molecular clock, likelihood ratio tests, relative-rate test.

Address for correspondence and reprints: David Posada, Variagenics, Inc., 60 Hampshire Street, Cambridge, Massachusetts 02139-1548. E-mail: dposada@variagenics.com.

*Mol. Biol. Evol.* 18(10):1976–1978. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

**Table 1**  
Probability of Rejecting the Molecular Clock

$\theta^a$	$\rho^b$	LRT		RRT
		Rejection at 5% Level	Average $\delta^c$	Rejection at 5% Level <sup>d</sup>
0.01 . . . .	0	0.068	9.428	0.021
	1	0.304	16	0.019
	4	0.562	24.974	0.022
	16	0.793	35.419	0.029
0.05 . . . .	64	0.909	41.016	0.043
	0	0.072	9.573	0.025
	1	0.454	33.561	0.029
	4	0.823	74.654	0.030
0.10 . . . .	16	0.975	132.837	0.029
	64	1.000	163.963	0.028
	0	0.075	9.725	0.023
	1	0.531	58.654	0.022
0.20 . . . .	4	0.912	134.921	0.021
	16	0.996	248.593	0.032
	64	1.000	291.537	0.025
	0	0.089	9.873	0.028
	1	0.595	116.768	0.021
	4	0.928	262.695	0.021
	16	1.000	448.143	0.035
	64	1.000	507.641	0.032

NOTE.—LRT = likelihood ratio test; RRT = relative-rate test.

<sup>a</sup>  $\theta$  is the expected mean pairwise difference among sequences per site.

<sup>b</sup>  $\rho$  is the number of recombination events in a gene per  $2N$  generations, where  $N$  is the effective diploid population size. The number of recombination events expected in the history of  $n$  sequences is given by  $\rho \sum_{i=1}^{n-1} 1/i$ . For the ingroup of 10 sequences, values of  $\rho$  of 0, 1, 4, 16, and 64 correspond, respectively, to 0, 2.83, 11.32, 45.26, and 181.05 recombination events.

<sup>c</sup>  $\delta$  is twice the difference in log likelihood between the clocklike model and the unconstrained rates model. For a  $\chi^2$  with 9 df,  $\delta$  is expected to be 9.

<sup>d</sup> After Bonferroni correction (family alpha level set to 5%).

most likely absent, and again we expected the same answer from the LRT and RRT. This was what we obtained, but in this case, the molecular clock was rejected, suggesting that recombination is probably absent and that the data do not fit a molecular clock.

As noted above, the relative ratio tests have been shown to lack statistical power in some cases (Bromham et al. 2000). This is consistent with the observation in the simulations that the type I error rate was around 7% for the LRT versus 3% for the RRT. Would the conclusions derived from the analysis of the empirical data sets still be the same if we make the RRT less conservative? The answer is yes. To obtain a 7% type I error rate for the RRT, the individual rejection level should be around 0.5%. When the RRT was applied with this alpha level to the four empirical data sets, the conclusions remained the same. In fact, the only difference was that in the HIV-1 data set, one more significant test was detected (21 instead of 20). Furthermore, if we set the RRT individual rejection level to 1% (equivalent to a 15% type I error rate), we still obtained the same results as before. Therefore, different powers between the LRT and the RRT do not seem to influence the conclusions of the empirical data analysis.

The RRT as described here is a conservative method for testing the molecular-clock hypothesis, independent of recombination. This is true only if the outgroup used did not recombine with the ingroup. There are two main applications of the RRT test:

**Table 2**  
Analysis of Empirical Data Sets

No. OF TAXA	OUTGROUP	LRT		RRT		CONCLUSION
		Clock -ln	Nonclock -ln	P	Significant Comparisons <sup>a</sup>	
10	<i>Vibrio</i> sp.	3,517.81	3,449.04	<0.0001	0/36	Recombination; Molecular clock
13	<i>Colobosaura modesta</i>	4,696.41	4,688.69	0.1632	0/66	No recombination; Molecular clock
20	HIV-1 group O	25,875.79	25,848.23	<0.0001	20/171	Recombination? No molecular clock
5	Fish	5,782.78	5,772.37	0.0001	2/6	No recombination? No molecular clock

NOTE.—Outgroup taxa were selected to minimize chances of their having recombined with the ingroup. LRT = likelihood ratio test; RRT = relative-rate test.

<sup>a</sup> After Bonferroni correction (family alpha set to 5%).

1. *The RRT can be applied to test for the presence of a molecular clock when recombination is likely to be present.* Recognizing the presence or absence of a clock may be relevant to understanding what kind of molecular evolutionary processes are acting upon the gene(s) under study. Rate variation among lineages is the footprint of selection and can be indicative of species radiations or differential structural constraints. It should be acknowledged that recombination as defined here includes processes as diverse as crossover, gene conversion, hybridization, and lateral transfer. Indeed, selection and recombination may occur simultaneously, and the RRT might be a useful tool to tear them apart. In addition, the RRT may be used to accurately identify those particular taxa that do not fit a molecular clock and/or have recombined. One of the most common applications of the molecular clock is the estimation of divergence times. However, it should be kept in mind that the fact that recombination is present might invalidate many of the applications of clock-based methods, most of which tacitly assume a lack of recombination (i.e., a single tree).
2. *Comparison of results of the LRT and RRT tests might suggest the presence of recombination.* The fact that the LRT rejects the molecular clock while the RRT fails to reject it might be due solely to the presence of recombination, as exemplified by the *Neisseria argF* data set. When this is the case, specific methods for detection of the presence of recombination (Robertson 2001) may be used to confirm and characterize the potential recombination events. When both tests fail to reject the clock, recombination should be absent or infrequent. Again, recombination detection methods could be applied to confirm this result.

In summary, the RRT can be a useful tool for investigating several molecular evolutionary processes, such as recombination and selection. The RRT is easily implemented in the software HYPHY (Kosakovsky and Muse 2000).

### Acknowledgments

Mikkel Schierup suggested the use of the relative ratio test for recombinant data. This manuscript benefited from conversations with Mikkel Schierup, Andrew Rambaut, and Michael Worobey. Thanks to Eddie Holmes and two anonymous reviewers for useful sug-

gestions. This work was supported by a BYU Graduate Studies Award and by an NSF Doctoral Dissertation Improvement Grant (NSF DEB 0073154).

### LITERATURE CITED

- BROMHAM, L., D. PENNY, A. RAMBAUT, and M. D. HENDY. 2000. The power of the relative rates tests depends on the data. *J. Mol. Evol.* **50**:296–301.
- CUNNINGHAM, C. W. 1997. Can three incongruence tests predict when data should be combined? *Mol. Biol. Evol.* **14**:733–740.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368–376.
- GRASSLY, N. C., and E. C. HOLMES. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.* **14**:239–247.
- HOCHBERG, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**:800–802.
- HUDSON, R. R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**:183–201.
- JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 in H. M. MUNRO, ed. *Mammalian protein metabolism*. Academic Press, New York.
- KOSAKOVSKY, S. L., and S. V. MUSE. 2000. HYPHY: hypothesis testing using phylogenies. Beta 1.7. Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh.
- MUSE, S. V., and B. S. WEIR. 1992. Testing for equality of evolutionary rates. *Genetics* **132**:269–276.
- RICE, W. R. 1989. Analyzing tables of statistical tests. *Evolution* **43**:223–225.
- ROBERTSON, D. L. 2001. Links to recombinant sequence detection/analysis programs. [http://grinch.zoo.ox.ac.uk/RAP\\_links.html](http://grinch.zoo.ox.ac.uk/RAP_links.html).
- SCHIERUP, M. H., and J. HEIN. 2000a. Recombination and the molecular clock. *Mol. Biol. Evol.* **17**:1578–1579.
- . 2000b. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**:879–891.
- SWOFFORD, D. L. 1998. PAUP\*: phylogenetic analysis using parsimony (\*and other methods). Version 4.0 beta. Sinauer, Sunderland, Mass.
- YANG, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**:1396–1401.
- ZHOU, J., and B. G. SPRATT. 1992. Sequence diversity within the *argf*, *fbp* and *reca* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Mol. Microbiol.* **23**:2135–2146.

EDWARD HOLMES, reviewing editor

Accepted June 20, 2001